Evaluating BERTopic on Open-Ended Data: A Case Study with Flemish Daily Narratives

Anonymous ACL submission

Abstract

Standard topic models often struggle to capture culturally specific nuances in text. This study evaluates the effectiveness of contextual embeddings for identifying culturally resonant themes in an underrepresented linguistic context. We compare the performance of KMeans Clustering, Latent Dirichlet Allocation (LDA), and BERTopic on a corpus of nearly 25,000 daily personal narratives written in Belgian-Dutch (Flemish). While LDA achieves strong performance on automated coherence metrics, subsequent human evaluation reveals that BERTopic consistently identifies the most coherent and culturally relevant topics, highlighting the limitations of purely statistical methods on this narrative-rich data. Furthermore, the diminished performance of K-Means compared to prior work on similar Dutch corpora underscores the unique linguistic challenges posed by personal narrative analysis. Our findings demonstrate the critical role of contextual embeddings in robust topic modeling and emphasize the need for human-centered evaluation, particularly when working with low-resource languages and culturally specific domains.

1 Introduction

Topic modeling is a cornerstone of text mining, facilitating the unsupervised discovery of latent thematic structures in large corpora. While traditional probabilistic models such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003) have been foundational, their underlying bag-of-words assumption limits their ability to capture complex semantics. A recent paradigm shift towards models like BERTopic (Grootendorst, 2022), a state-of-the-art (SOTA) model which leverages contextualized embeddings from pre-trained transformers, has shown significant promise in generating more semantically coherent topics. These models can capture nuanced relationships, including domain-specific named entities and morphologically rich constructs, critical for linguistically complex data.

However, despite this progress, two significant gaps persist in the literature. First, research has overwhelmingly focused on high-resource, standardized languages, with a lot of scope left for under-resourced languages to be unexplored. This focus not only limits the generalizability of existing models but also risks perpetuating a technological bias where the nuances of smaller linguistic communities are overlooked. Models trained on standard corpora often fail to capture the unique lexical and semantic patterns of regional dialects or sociolects, leading to a superficial or even inaccurate understanding of the underlying discourse (Kamiloğlu et al., 2025). Second, the predominant application domain has been structured or short-form text like news articles or social media posts (Egger and Yu, 2022; Schäfer et al., 2024), while the challenges of modeling unstructured, open-ended personal narratives have received less attention. Distinct from the short-form, often decontextualized nature of social media data, daily narratives provide granular, contextually-grounded accounts of lived experience. These texts capture a wide spectrum of daily activities, affective states, and cognitive processes, thereby constituting a rich and challenging corpus for studying naturalistic language use.

This study addresses these gaps through a systematic evaluation of BERTopic on a novel corpus of informal, medium-length Flemish daily narratives that reflect diverse lived experiences in Belgium. This data presents a unique challenge due to its inherent linguistic variability, fluid thematic structure, and use of highly contextual, culturally specific language. We benchmark BERTopic's performance against two established baselines: the probabilistic LDA model (Blei et al., 2003; Shin et al., 2025) and a Term Frequency-Inverse Document Frequency (TF-IDF) (Salton and Buckley,

1988; Ramos, 2003) based KMeans clustering approach (MacQueen, 1967; Lloyd, 1982; Berkhin, 2006; Sinaga and Yang, 2020). Our analysis moves beyond a sole reliance on automated coherence metrics, which we argue can be insufficient or even misleading for embedding-based models for this type of data. Instead, we employ a hybrid evaluation framework that integrates these metrics with human evaluation of topic interpretability and relevance. We demonstrate that for this task, the contextual awareness of BERTopic yields qualitatively superior topics, highlighting the need to advance NLP methodologies for underrepresented languages and advocating for more nuanced evaluation protocols.

The remainder of this paper is structured as follows: Section 2 provides an overview of KMeans, LDA, and BERTopic. Section 3 discusses the methodology (including data collection and data analysis) employed for this study. Section 4 presents the results. Section 5 discusses the findings, section 6 presents the limitations, and section 7 concludes the study.

2 An Overview of The Three Topic Modeling Approaches: KMeans, LDA, and BERTopic

In this section, we compare and detail the key components of the specific topic modeling approaches chosen for this study: a TF-IDF based KMeans clustering baseline, the probabilistic Latent Dirichlet Allocation (LDA), and the embedding-based BERTopic.

2.1 KMeans

To represent a non-probabilistic, geometric clustering baseline, we employ KMeans (MacQueen, 1967; Lloyd, 1982; Berkhin, 2006; Sinaga and Yang, 2020). For vectorization, we use TF-IDF, as it is a standard feature extraction method for classical document clustering and provides a stronger baseline than raw term counts due to its ability to down-weight frequent, uninformative words (Salton and Buckley, 1988; Ramos, 2003). Following prior work demonstrating its efficacy on Dutch texts (Kamiloğlu et al., 2025), we apply KMeans to partition the resulting TF-IDF document vectors, where each cluster centroid represents a latent topic.

We deliberately avoided using embedding-based vectorizers for this baseline to ensure a clear comparison between distinct modeling paradigms. The objective of this study is to contrast an end-to-end, embedding-native model (BERTopic) against established, non-embedding methods. Using embeddings to generate features for KMeans would position it as a hybrid model, which, while a valid approach, falls outside the scope of this direct comparison.

2.2 Latent Dirichlet Allocation (LDA)

As a canonical probabilistic baseline, we include LDA, a traditional probabilistic topic modeling technique that assumes a hierarchical Bayesian structure, where documents are generated by sampling topics from a Dirichlet-distributed prior, with words subsequently sampled from topic-specific multinomial distributions (Blei et al., 2003; Shin et al., 2025). It employs CountVectorizer (Pedregosa et al., 2011) to transform preprocessed text into a document-term matrix (DTM), where each row represents a document, and each column corresponds to a unique term.

2.3 BERTopic

BERTopic is a relatively recent topic modeling approach that leverages transformer-based embeddings to capture nuanced semantic relationships in textual data and generate topics (Grootendorst, 2022). The modular, multi-stage pipeline primarily involves three steps: dimensionality reduction, document clustering, and topic extraction.

2.3.1 Dimensionality Reduction

To address the challenges posed by the high dimensionality of the embeddings, we employ Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) to reduce the high dimensional embeddings to a low dimensional space. UMAP is a non-linear technique adept at preserving both the local and global structure of the data from the high-dimensional space, which is crucial for effective subsequent clustering (Grootendorst, 2022).

2.3.2 Clustering

The reduced embeddings are then clustered using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) (Campello et al., 2013). We selected HDBSCAN as the default clustering component of BERTopic for several reasons that make it particularly well-suited for our corpus of open-ended personal narratives.

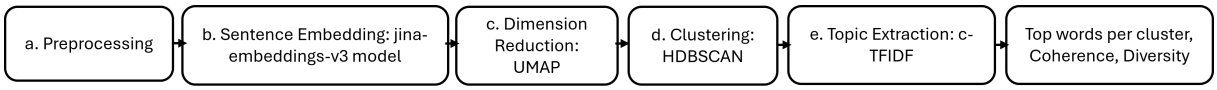

Figure 1: BERTopic implementation pipeline

First, unlike centroid-based algorithms such as KMeans, HDBSCAN does not require the number of topics to be specified *a priori*. This is a significant advantage when exploring a novel dataset where the thematic structure is unknown. Second, HDBSCAN is a density-based algorithm that identifies clusters of varying shapes and densities, which is a better fit for the fluid and unstructured nature of narrative text than the globular cluster assumption of KMeans.

More crucially, HDBSCAN can classify data points as noise (outliers) rather than forcing every document into a topic. This is invaluable for our dataset, as personal narratives can often be non-topical or too ambiguous to be assigned to a coherent theme. By treating these as noise, we ensure that the resulting topic representations are cleaner and more semantically cohesive. While BERTopic's modularity allows for other clustering algorithms, our use of HDBSCAN is a deliberate choice to align the modeling approach with the inherent characteristics of our data (Grootendorst, 2022).

2.3.3 Topic Extraction

To provide interpretable representations of each topic, the key terms are extracted using BM25-weighted class-based TF-IDF (c-TF-IDF) (Grootendorst, 2022). This method treats all documents within a given cluster as a single composite document. It then calculates term frequencies within this composite document and weights them by the inverse frequency of the term across all other clusters. This process extracts words that are most representative of a specific topic cluster.

3 Methodology

We conduct a comparative analysis of KMeans, LDA, and BERTopic on a novel corpus of Flemish daily narratives. Our evaluation framework is twofold, comprising a suite of automated topic coherence and diversity metrics, supplemented by human evaluation to assess topic interpretability.

3.1 Data and Data Collection

The dataset consists of open-ended responses of daily narratives in Flemish with a total of 24,854 texts. Participants $(N = 115, age : 18 - 65, M = 26.47, SD = 8.87)$ were recruited via flyers, online posts, and word-of-mouth. Eligible individuals were native Dutch speakers living in Belgium, had to be at least 18 years old, and had to be in possession of a working smartphone. Over 70 days, participants received four prompts per day via a smartphone app called m-Path (Mestdagh et al., 2023), responding to the question: "What is going on now or since the last prompt, and how do you feel about it?" They provided either a one-minute voice message or a 3-4 sentence text response. For additional details on participant recruitment, study procedures, data collection and preparation see Appendix.

3.2 Data Analysis

3.2.1 Preprocessing

The raw texts underwent a multi-step preprocessing pipeline: (i) removal of dataset-specific annotation tags, and author-identifying references (ii) lemmatization using the Stanza (Qi et al., 2020) model for Dutch, and (iii) filtering of documents shorter than 15 words. A custom, corpus-specific stopword list was also applied. The final processed corpus contains 24,752 documents.

3.2.2 Embedding Model Selection for BERTopic

To identify the optimal sentence-level embeddings for Flemish, we evaluated three multilingual models: (a) *robert-2022-dutch-sentence-transformers* (Netherlands Forensic Institute, 2024): A Dutch-specific model fine-tuned for semantic similarity task (b) *gte-multilingual-base* (Zhang et al., 2024): A general-purpose multilingual encoder optimized for long-context retrieval (c) *jina-embeddings-v3* (Sturua et al., 2024) : A multilingual model leveraging Task LoRA for adaptable semantic representation. Qualitative assessment of the resulting topics indicated that *jina-embeddings-v3* produced the most coherent themes for our specific corpus. We attribute this to its XLM-

ROBERTa architecture augmented with Rotary Position Embeddings and Task LoRA adapters. Consequently, all BERTopic experiments utilize this model, which generates 1,024-dimensional embeddings via the SentenceTransformers library (Reimers and Gurevych, 2019). We then utilized UMAP to reduce these embeddings to a 5-dimensional space.

3.2.3 Model Configuration and Hyperparameter Optimization

KMeans Clustering

For the KMeans baseline, documents were vectorized using TF-IDF, a classical, non-semantic feature space following the approach of (Kamiloğlu et al., 2025). To establish a robust baseline, we conducted a systematic exploration of vectorizer parameters, testing various combinations of min_df and max_df, including both absolute counts and proportional thresholds. The final configuration reported in this study (min_df=0.01, max_df=0.9) was selected to generate a compact vocabulary focused on high-frequency terms, resulting in a DTM of size 24,725 $\times$ 326. Sublinear term frequency scaling (Manning et al., 2008) was also applied to reduce the influence of high-frequency terms. For the clustering step, the optimal number of clusters (k) was determined via Silhouette analysis (Rousseeuw, 1987) over a range of $k = 2$ to 200, which indicated a peak score at $k = 128$ (see Appendix). The entire pipeline was implemented using Python's scikit-learn (Pedregosa et al., 2011) library with fixed random seeds to ensure reproducibility.

LDA

For the LDA baseline, we used the scikit-learn implementation (Pedregosa et al., 2011). We tuned the number of topics(k)(ranging from 5 to 150), α (`doc_topic_prior`), and β (`topic_word_prior`) over [0.01, 0.1, 1.0, 0.5]. Model selection was guided by coherence scores computed with the Gensim coherence model (Rehurek and Sojka, 2011).

BERTopic

We systematically tuned BERTopic's hyperparameters using *jina-embeddings-v3* to optimize topic coherence and diversity. For dimensionality reduction, we tested `umap_n_neighbors` with values [10, 15, 20, 25, 30] and `umap_min_dist` with values [0.0, 0.1]. For clustering, we evaluated `hdbscan_min_cluster_size` with values [5, 10, 15]. For the model, we tested `min_topic_size` with values [5, 10, 15] and `nr_topics` with the value [`auto`]. Initial clustering with BERTopic labeled 59% of documents (n = 14,600) as outliers. To mitigate this, we employed the *Approximate Distribution Method*, which probabilistically assigns outliers to the nearest topic based on cosine similarity in embedding space (Grootendorst, 2022). This reduced outliers to 7.56% (n = 1,872) for the best BERTopic C_v score observed.

3.3 Evaluation Metrics

To provide a comprehensive comparison, we evaluate the models on both automated metrics and human judgments.

3.3.1 Automated Metrics

We employ a suite of four standard topic coherence metrics (Mimno et al., 2011) to evaluate the interpretability and semantic consistency of topics generated by BERTopic and LDA. C_v measures coherence based on a combination of Normalized Pointwise Mutual Information (NPMI) and cosine similarity over a sliding window (size 110), closely aligning with human judgments of word co-occurrence (Röder et al., 2015). C_{npmi} is a normalized PMI score robust to frequency biases, scaled between -1 (dissociation) and 1 (perfect association), making it suitable for comparing embedding-based models (where semantic similarity extends beyond raw co-occurrence), and count-based models. As a normalized and stable variant of PMI, it is less susceptible to biases from word frequency, making it a valuable metric for comparing embedding-driven models like BERTopic (which prioritize semantic similarity over raw co-occurrence) against count-based models like LDA. U_{mass} evaluates topic quality based on the log conditional probability of word co-occurrences, favoring probabilistic models like LDA. Finally, U_{uci} uses raw PMI as a baseline co-occurrence measure. We also compute topic diversity, defined as the percentage of unique words in the top-10 words of all topics.

3.3.2 Human Evaluation of the Topics

To complement the automated coherence metrics, we conducted a human evaluation study to assess the semantic coherence of the topics generated by each model. Two human annotators, both Dutch speakers, participated in a word intrusion task. This

task is used to probe whether topic words form a semantically cohesive group by asking annotators to identify a word that does not semantically belong (Chang et al., 2009; Bhatia et al., 2018), and is a widely used method for evaluating topic coherence.

For each of the three models (KMeans, LDA, BERTopic), we randomly sampled 40 topics. For each topic, we constructed a six-word set comprising the top five words associated with the topic and a sixth "intruder" word. Intruder words were drawn from high-probability terms of unrelated topics within the same model. Specifically, an intruder was sampled with uniform probability from the top 10 words of a randomly selected, non-related topic. This method ensures that the intruder is a plausible, high-frequency word in its own right, making the task a robust test of the target topic's semantic coherence. The order of the sixth (intruder) word was randomized for each trial to minimize position bias. Annotators were instructed to independently identify the single word in each set that did not semantically fit with the others. We report two evaluation metrics in this regard: (i) Topic Coherence Accuracy: This constitutes the proportion of topic sets in which the annotator correctly identified the intruder word. This serves as an estimate of the interpretability and semantic coherence of the topics (Chang et al., 2009) (ii) Inter Annotator Agreement (IAA): To assess annotation reliability, we computed Krippendorff's Alpha α (Krippendorff, 2004; Artstein and Poesio, 2008), a standard metric for agreement on categorical judgments. Here, agreement reflects whether annotators identified the same intruder word in each set. High agreement indicates consistent judgments about topic coherence across annotators. Figures illustrating annotator accuracy and inter-annotator agreement are provided in the appendix.

4 Results and Analysis

Our analysis compares the embedding-based BERTopic model against a probabilistic LDA baseline and a TF-IDF-based KMeans baseline on a corpus of over 24,000 Flemish daily narratives. The results reveal a significant divergence between automated coherence metrics and human-perceived topic quality, highlighting the limitations of traditional evaluation paradigms for semantically-aware models on linguistically complex data.

4.1 Quantitative Analysis

The quantitative evaluation, summarized in Tables 1 and 2, presents a nuanced picture. In line with prior work on topic models applied to short or medium length texts, performance of both BERTopic and LDA peaked at an optimal number of topics before declining, particularly in datasets with limited word counts per document (Aggarwal and Zhai, 2012; Muthusami et al., 2024; Mutsaddi et al., 2025).

When comparing the best-performing configurations of each model, we found that LDA produced higher C_v topic diversity scores than BERTopic (Table 2). Specifically, LDA achieved a C_v score of 0.5430 for 120 topics, and BERTopic achieved 0.3412 for 76 topics. The higher score for LDA suggests stronger word-level topical coherence, likely due to its reliance on explicit term co-occurrence statistics, which aligns with traditional interpretability expectations (Röder et al., 2015). On the other hand, the lower C_v score of BERTopic could reflect its dependency on contextual embeddings, which prioritize semantic similarity over lexical overlap, a known limitation of PMI-based metrics in embedding-driven models (Grootendorst, 2022). Table 1 presents the C_v scores for selected topic counts with both models. Topic counts with very poor coherence scores have been omitted.

Model	No. of Topics	C_v Score
BERTopic	60	0.2966
	76	**0.3412**
	110	0.3370
	117	0.3379
LDA	76	0.519
	100	0.505
	120	**0.543**
	148	0.542

Table 1: C_v scores for selected topic counts with BERTopic and LDA.

However, BERTopic produced higher scores on the other coherence metrics (C_{npmi}, U_{mass}, U_{uci}) than LDA, suggesting that C_{npmi}'s stability accommodated BERTopic's embedding-based semantics better than raw PMI (Lau et al., 2014). Furthermore, U_{mass}, which penalizes rare word pairs, may align better with BERTopic's ability to cluster medium-length texts without overfitting to sparse co-occurrences. Table 2 compares the scores of the best BERTopic and best LDA models across four

coherence metrics and topic diversity.

In terms of topic diversity, LDA achieved a higher score of 0.9675 for 120 topics, compared to BERTopic's 0.8455. We interpret these findings to indicate that LDA's Dirichlet prior encourages distinct topic distributions, while BERTopic's HDBSCAN clustering allows overlapping semantic themes, which could reduce diversity but potentially capture nuanced relationships.

The KMeans baseline, optimized using geometric metrics like the Silhouette score, favored a high number of clusters (k=128) (see Appendix). However, this granularity resulted in a marked decline in topic coherence and interpretability, underscoring the potential mismatch between geometric partitioning objectives and the goal of extracting semantically meaningful themes (Aggarwal and Zhai, 2012).

4.2 Qualitative Analysis: Best KMeans vs. Best LDA vs. Best BERTopic

While automated metrics like C_v favour LDA, human evaluation revealed significant discrepancies in semantic coherence and contextual relevance, particularly evident in the morphologically rich and regionally specific Flemish corpus. BERTopic consistently generated thematically cohesive and culturally resonant topics, while LDA and KMeans struggled with semantic fragmentation and noise. The themes identified by BERTopic were specific and interpretable, for example differentiating everyday routine ('workday routine', 'planning and communication', and 'studying') from activities ('horse riding', 'travel and outdoor recreation', and 'film evenings'), "Chiro" (Belgium's largest youth organization), "rains" (as it frequently rains in Belgium) and overall emotional and mental state ('headaches and migraine-related pain', 'academic stress and assignments').

BERTopic's superior ability to capture relevant themes can be illustrated with the topics related to "studying and academic life" (Table 3). It successfully captures "*aula_vriend*" (*lecture hall friend*) and "*bibliotheek*" (*library*), while the LDA model conflated academic terms with noisy co-occurrences like "*bus*" and "*toekomst*" (*future*), likely due to its reliance on document-level word distributions, which struggle with sparse co-occurrence patterns in medium-length texts. KMeans clustered high-frequency but semantically unrelated verbs "*eten*" (*to eat*) and "*slapen*" (*to sleep*), failing to isolate domain-specific themes.

Model	Topic No.	Topic Words
BERTopic	7	**studeren**, bibliotheek, tevreden, aula_vriend, studie_tijd, leeszaal, bib, dezeochtend, basically
LDA	108	**studeren**, bus, snappen, interessant, deadline, toekomst, af_sluiten, practicum, focus
KMeans	73	**studeren**, goed, vandaag, eten, beginnen, rest, dag, ver, moe, tijd, oké, weten, slapen, morgen, zin

Table 3: Topics identified by BERTopic, LDA, and KMeans related to "studying and academic life"

This pattern persisted in domains like "fitness" (Table 4). BERTopic cohesively grouped terms such as "*fitnessen*" (*to work out*), "*gym*," and "*kracht_training*" (*power training*), while LDA's topic included semantically discordant words such as "*broer*" (*brother*) and "*baby*." KMeans conflated fitness with unrelated daily activities such as "*eten*" (*to eat*) and "*moe*" (*tired*). Another trend observed in our KMeans analysis is that many common words (underlined in the tables) reappeared across clusters.

Model	Topic No.	Top Words
BERTopic	4	**fitnessen**, fitness, workout, gym, oefening, trainen, joggen, kracht_training, sport_les, sport_school
LDA	48	**fitnessen**, time, baby, start, hasten, toe_voegen, quality, model, aanpassing
KMeans	118	**fitnessen**, goed, eten, dag, vandaag, weten, leuk, zin, vanavond, moe, beginnen, studeren, rest, werken, proberen

Table 4: Topics identified by BERTopic, LDA, and KMeans related to "fitness"

4.3 Human Evaluation Results

Our evaluation showed that BERTOPIC achieved notably higher precision (Annotator 1: 95.0 %, Annotator 2: 87.5 %) compared to KMeans (Annotator 1: 60.0 %, Annotator 2: 52.5.0 %) and LDA (Annotator 1: 20.0%, Annotator 2: 25.0%). This substantial gap indicated that topics produced by BERTopic have better interpretability.

To assess the consistency and ensure the reliability of annotators' judgments, we also computed the Inter-Annotator Agreement (IAA) using Krippendorff's Alpha (Krippendorff, 2004). Agreement was high for BERTOPIC ($\alpha = 0.874$), further

Model	No. of Topics	C_v	C_{npmi}	U_{mass}	U_{uci}	Topic Diversity
BERTopic	76	0.3412	**-0.1619**	**-12.2167**	**-5.6599**	0.8455
LDA	120	**0.543**	-0.21	-16.39	-5.93	**0.9675**

Table 2: Topic coherence and topic diversity for the best hyperparameter configurations.

reinforcing confidence in the model's topic coherence. In contrast, moderate agreement levels were observed for KMeans ($\alpha = 0.545$) and LDA ($\alpha = 0.547$), suggesting relatively lower coherence and greater ambiguity in topics generated by these models. These results highlight BERTopic's strength in producing semantically coherent and intrepretable topics.

5 Discussion

This study's findings reveal a critical tension in topic model evaluation: the divergence between automated coherence metrics and human judgments of topic quality, particularly when applied to linguistically complex corpus of Flemish daily narratives.

The superior C_v score achieved by LDA, juxtaposed with its poorer performance in human evaluation, points to a systemic limitation of metrics that rely on surface-level co-occurrence. As evidenced by the qualitative analysis, LDA's inclusion of frequent but semantically irrelevant co-occurrences (e.g., "*bus*" near "*studeren*" [*to study*]) could reflect its bias toward "syntagmatic associations" (statistical proximity) rather than true "paradigmatic relevance" (thematic consistency). In contrast, BERTopic's embeddings prioritize contextual relationships (e.g., "*bibliotheek*" [*library*] ↔ "*studie_tijd*" [*study time*]), aligning with human intuition (Lau et al., 2014).

This discrepancy strongly suggests that an over-reliance on automated metrics like C_v can be misleading for evaluating embedding-based models, whose strengths lie in capturing semantic nuances that bag-of-words approaches miss. Furthermore, the high inter-annotator agreement for BERTopic (Krippendorff's α = 0.874), compared to the moderate agreement for the baselines, indicates that BERTopic's topics are not only more coherent but also more clearly and consistently interpretable. Our results thus underscore the necessity of adopting hybrid evaluation frameworks that integrate human-in-the-loop validation, especially for under-resourced languages and non-standard text genres (Muthusami et al., 2024).

Finally, our results with the KMeans baseline warrant discussion. While prior work demonstrated the efficacy of KMeans on Dutch texts (Kamiloğlu et al., 2025), our experiments did not replicate this success, instead yielding generic and incoherent clusters. We posit that this is not a failure of the algorithm itself, but rather a limitation of the TF-IDF feature space when applied to the unstructured, highly variable style of personal narratives. The vocabulary pruning inherent in TF-IDF, even with systematic tuning, appears insufficient to create a vector space where thematically distinct narratives form separable geometric clusters, resulting in the observed low-quality output.

6 Limitations

Our study, while providing a focused analysis of topic modeling on a unique dataset, is subject to several limitations that frame the scope of our conclusions.

First, the generalizability of our findings is constrained by our use of a single, highly specific corpus: daily personal narratives from a predominantly young adult, Dutch-speaking population. The linguistic characteristics of this genre: informal, unstructured, and reflective differ significantly from other text types like news articles or formal documents. Consequently, the relative performance of BERTopic and the baseline models may not directly translate to other domains, languages, or demographic groups. The topics themselves, such as "studying" and "internships," reflect the life stage of our participants and are not representative of the entire population.

Second, our methodology is dependent on the quality of upstream NLP components, which have known limitations for non-standard language variants. We selected the *jina-embeddings-v3* model after a qualitative comparison, but even state-of-the-art multilingual models may not optimally capture the dialect-specific nuances of Flemish compared to a hypothetical native model. Furthermore, our preprocessing pipeline relied on Stanza for lemmatization. Errors or inconsistencies in lemmatizing dialectal or informal terms (e.g. *dezeochtend*) can

introduce noise that disproportionately affects the performance of count-based models like LDA and the TF-IDF-based KMeans, potentially impacting the fairness of our comparison.

Finally, while this study advocates for the superiority of BERTopic's qualitative output, we acknowledge the inherent subjectivity of topic model evaluation. Our conclusions are based on a combination of automated metrics and human judgment, yet both have constraints. Automated coherence scores are imperfect proxies for human interpretation, and human evaluation, while essential, is difficult to scale and can be influenced by the specific task design and the evaluators' own biases.

7 Conclusion and Future Work

This study evaluated the efficacy of BERTopic for modeling unstructured, open-ended daily narratives in Flemish. Our findings demonstrate that while traditional models like LDA can achieve higher scores on co-occurrence-based metrics such as C_v, BERTopic generates qualitatively superior topics that are more semantically coherent and culturally resonant. This work highlights a critical challenge in the field: the potential for automated metrics to be misleading when evaluating modern, embedding-based models on morphologically rich and context-dependent text, underscoring the need for hybrid evaluation frameworks.

The validation of BERTopic for extracting meaningful themes from noisy, real-world narratives provides a crucial methodological foundation for downstream applications in computational social science. Traditional text analysis in these domains often relies on lexicon-based tools like LIWC (Pennebaker et al., 2001), which are limited by their static, non-contextual nature. Our work shows that modern topic models can offer a more nuanced, data-driven lens into the thematic content of personal experiences. This research direction promises to yield more robust and context-aware tools for the computational analysis of human experience.

Ethical Considerations

All participants provided informed consent and were fully briefed on the study's procedures, including the secure collection of self-reported responses and sensor data. Personal and sensitive information was anonymized and handled with the highest standards of confidentiality. Compensation was provided uniformly according to the study protocol, ensuring fair treatment throughout the 70-day experience sampling process.

Data Availability Statement

The dataset of Flemish daily narratives used in this study contains sensitive personal information and was collected from participants in Europe under the General Data Protection Regulation (GDPR). To protect participant privacy while supporting scientific reproducibility, the anonymized dataset will be made available to qualified researchers for non-commercial research purposes. Access can be requested by contacting the corresponding author and requires the signing of a data use agreement, which will stipulate that the work must be cited in any resulting publications. The code used to conduct the experiments reported in this paper, including preprocessing scripts and model implementations, will be made publicly available on a GitHub repository upon publication.

References

Charu C. Aggarwal and ChengXiang Zhai. 2012. A survey of text classification algorithms. In Charu C. Aggarwal and ChengXiang Zhai, editors, *Mining Text Data*, pages 163–222. Springer, Boston, MA.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Pavel Berkhin. 2006. A survey of clustering data mining techniques. In Jacob Kogan, Charles Nicholas, and Marc Teboulle, editors, *Grouping Multidimensional Data*, pages 25–71. Springer.

Sumit Bhatia, Jey Han Lau, and Timothy Baldwin. 2018. Topic intrusion for automatic topic model evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2163–2173, Brussels, Belgium. Association for Computational Linguistics.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(1):993–1022.

Ricardo J. G. B. Campello, Davoud Moulavi, and Jörg Sander. 2013. Density-based clustering based on hierarchical density estimates. In Jian Pei, Vincent S. Tseng, Longbing Cao, Hiroshi Motoda, and Guanfeng Xu, editors, *Advances in Knowledge Discovery and Data Mining. PAKDD 2013. Lecture Notes in Computer Science (Vol. 7819)*, pages 160–172. Springer, Berlin, Heidelberg.

Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. Reading

tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems 22 (NeurIPS)*.

Roman Egger and Joanne Yu. 2022. A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts. *Frontiers in Sociology*, 7:886498.

Maarten R. Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Roza G. Kamiloğlu et al. 2025. What makes us feel good? a data-driven investigation of positive emotion experience. *Emotion*, 25(1):271–276.

Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology*, 2 edition. Sage Publications.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 530–539, Gothenburg, Sweden. Association for Computational Linguistics.

Stuart P. Lloyd. 1982. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137.

J. MacQueen. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press.

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861.

Maarten Mestdagh, Stijn Verdonck, Maarten Piot, Kris Niemeijer, Ghina Kilani, Francis Tuerlinckx, Peter Kuppens, and Eline Dejonckheere. 2023. m-path: An easy-to-use and highly tailorable platform for ecological momentary assessment and intervention in behavioral research and clinical practice. *Frontiers in Digital Health*, 5.

David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 262–272. Association for Computational Linguistics.

R. Muthusami, N. Mani Kandan, K. Saritha, et al. 2024. Investigating topic modeling techniques through evaluation of topics discovered in short texts data across diverse domains. *Scientific Reports*, 14:12003.

Atharva Mutsaddi, Anvi Jamkhande, Aryan Shirish Thakre, and Yashodhara Haribhakta. 2025. BERTopic for topic modeling of hindi short texts: A comparative study. In *Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages*, pages 22–32, Abu Dhabi. Association for Computational Linguistics.

Netherlands Forensic Institute. 2024. robbert-2022-dutch-sentence-transformers (revision cdf42f6).

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

James W. Pennebaker, Martha E. Francis, and Roger J. Booth. 2001. *Linguistic Inquiry and Word Count: LIWC 2001*. Lawrence Erlbaum Associates, Mahwah, NJ.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108. Association for Computational Linguistics.

Juan E. Ramos. 2003. Using tf-idf to determine word relevance in document queries. Technical report.

Radim Rehurek and Petr Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2).

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992. Association for Computational Linguistics.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM)*, pages 399–408. Association for Computing Machinery.

Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Gerard Salton and Chris Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing Management*, 24(5):513–523.

Karla Schäfer, Jeong-Eun Choi, Inna Vogel, and Martin Steinebach. 2024. Unveiling the potential of BERTopic for multilingual fake news analysis – use case: Covid-19.

Eunyong Shin, Sun Yim, and A Ra Koh. 2025. Comparison of consumer perceptions of sustainable and ethical fashions pre- and post-COVID-19 using LDA topic modeling. *Humanities and Social Sciences Communications*, 12(1):226.

Kusuma Prayoga Sinaga and Ming-Shing Yang. 2020. Unsupervised k-means clustering algorithm. *IEEE Access*, 8:80716–80727.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Andreas Koukounas, Nan Wang, and Han Xiao. 2024. jina-embeddings-v3: Multilingual embeddings with task LoRA. *Preprint*, arXiv:2409.10173.

Bastian Tamm, Jordi Poncelet, Manon Barberis, and Marjolein Vandermosten. 2024. Weakly supervised training improves flemish asr of non-standard speech. In *Dutch Speech Tech Day 2024*, Hilversum, The Netherlands.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, et al. 2024. mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412.

Appendix

A Participants

Participants were a community sample recruited through flyers, online posts, and word-of-mouth. To be eligible, participants needed to be native Dutch speakers at least 18 years old, living in Belgium, with a smartphone in good condition. Interested individuals were directed to an eligibility survey, and these criteria were later verified during the online introduction session.

Participants were compensated up to €250 for completing a 70-day (10-week) experience sampling protocol, including biweekly online surveys. They earned €0.50 for each completed experience sampling prompt (4 prompts $\times$ 70 days = 280 prompts or €140 max), €10 for each short online survey (2, 4, 6, and 8 weeks; €40 max), and €15 for each long online survey (0 and 10 weeks, €30 max). Participants who remained in the study for at least 60 days received a bonus of €40. Compensation was provided as a lump sum at the study's end. To remain in the study, participants needed to complete at least 75% of prompts and submit verbal descriptions of at least 25 words. Compliance was monitored via periodic checks and summary reports.

A total of 115 participants enrolled (age: 18–65, M = 27.26, SD = 9.86; gender: 58 women, 56 men, 1 other). Of these, 10 left the study early, and 3 were dismissed for low compliance (i.e., response rate under 50%). The remaining 102 participants (age: M = 26.47, SD = 8.87) included 52 women, 49 men, and 1 other.

Study procedures and materials were reviewed and approved by the KU Leuven Social and Societal Ethics Committee (SMEC), protocol G-2023-6379-R3(AMD). Data collection occurred from August 2023 through July 2024, with all study instructions provided in Dutch.

Procedure and Materials

Participants completed 70 days of experience sampling, receiving four prompts daily via a dedicated smartphone app (*m-Path*) (Mestdagh et al., 2023). Prompts were sent pseudo-randomly between 9 AM and 9 PM, with at least an hour between them. At each prompt, participants responded to: "What is going on now or since the last prompt, and how do you feel about it?" Responses were recorded as one-minute voice messages or 3–4 sentence typed responses.

While completing prompts, the *m-Path* app used phone sensors to record GPS coordinates, ambient noise levels, step count, and recent phone app usage (Android devices only, app names recorded). Participants also completed biweekly online surveys assessing well-being and emotional functioning, but these data were not used in the current analysis.

Data Preparation

Voice recordings were automatically transcribed using a proprietary algorithm developed at the KU Leuven Department of Electrical Engineering (ESAT) (Tamm et al., 2024). These transcriptions were merged into a master dataset, integrating both transcribed and typed responses for analysis.

Model evaluations were performed using an NVIDIA RTX-5000 GPU and implemented in PyTorch

B Silhouette Analysis

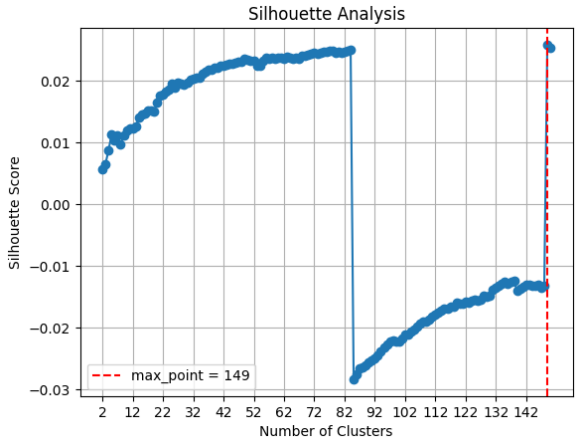

Figure 2: Silhouette Score vs Number of Clusters

C Human Annotation Results

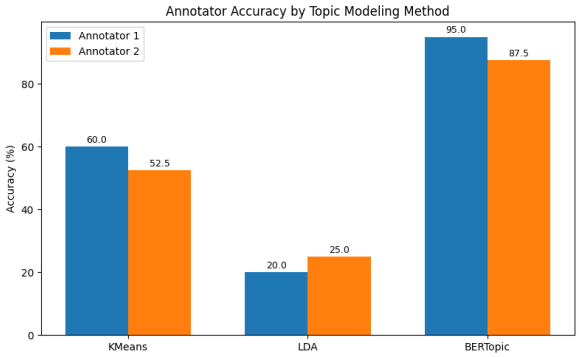

Figure 3: Annotator's Accuracy

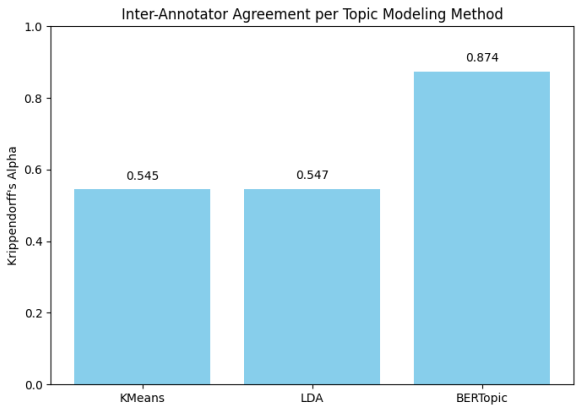

Figure 4: Inter-Annotator Agreement